Part II - Advanced Issues in Stream Classification

Gold Coast, 16th of April 2013

Georg Krempl and Myra Spiliopoulou
Knowledge Management & Discovery
Otto-von-Guericke University Magdeburg
georg.krempl@ovgu.de

Thanks to my colleagues at the KMD lab, Vera Hofer, Emilie Morvant, and to Indrė Zliobaitė.

Agenda

Focus of this part

- Advanced problems in classification (and regression) tasks,
- formulated and described using multiple streams.

Deliverables

- Framework that integrates and highlights several feedback issues.
- A brief overview on solutions and open questions for some feedback issues.

Outline

- Motivation and introduction
- Feedback issues as multiple streams problem
- Concept drift
- Verification latency
- Active learning

Classification in Static, Batch-Oriented Contexts

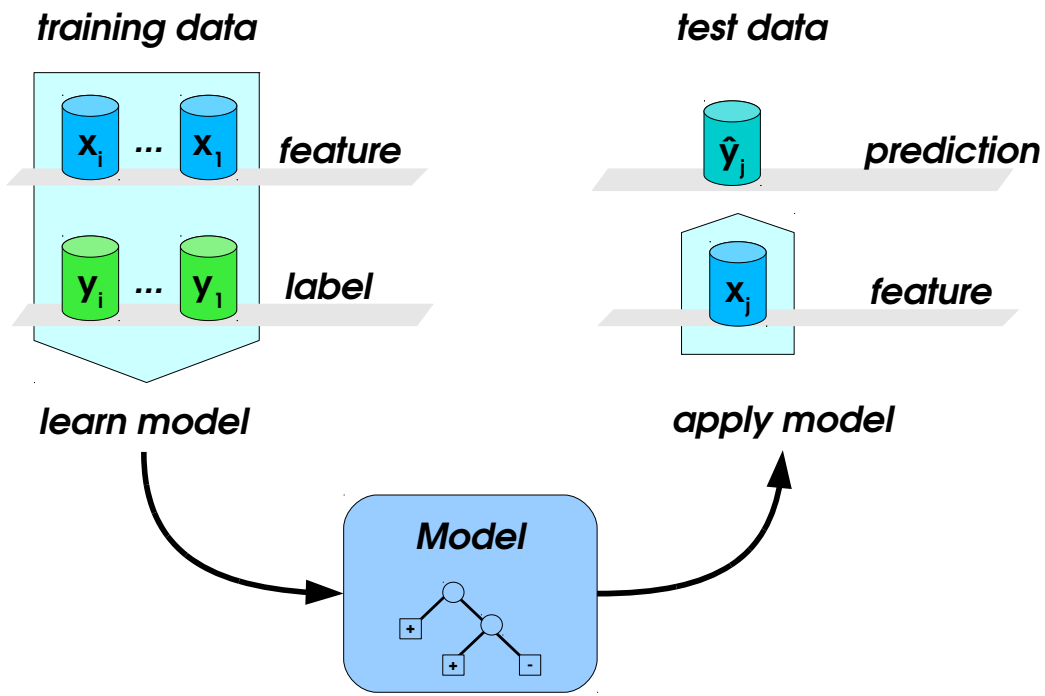


Figure: Typical Static Classification Model

Motivations for Stream Mining

Some applications like

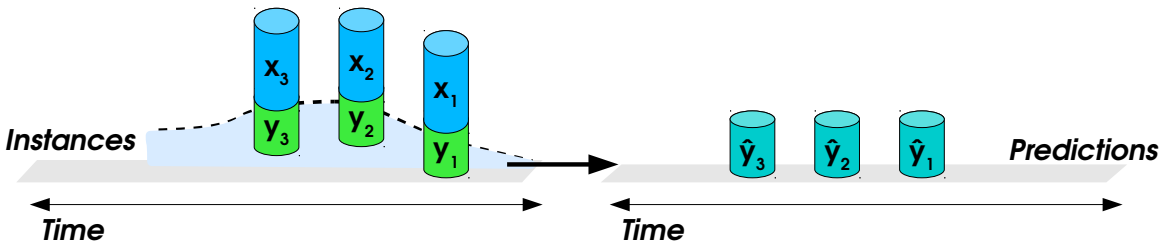
- sensor networks,
- financial data streams,
- web logs, ...

challenge the static, batch-oriented view:

- Data is received continuously
- Big (possibly infinite) data sets
- Prediction must be available 24/7
- Model should reflect current knowledge

- Problem is better formulated as *stream mining* task

Data Mining in Evolving Data Streams



Challenges

- Streams: How to handle small/big data?
  - Limited training data at the beginning,
  - increasing amount of (training) data over time
  - How to avoid increasing *space* and *time* complexity?
- Change over time: Use all data?
  What can change?
  - Feature space:
    Adding new variables,
    Removing old variables,
    Change in domain of variable
  - Label domain
  - Distributions: Change of $P(Y|X)$ due to *drift* of $P(X|Y)$, $P(Y)$, $P(X)$
    Use of old data may deteriorate model quality!

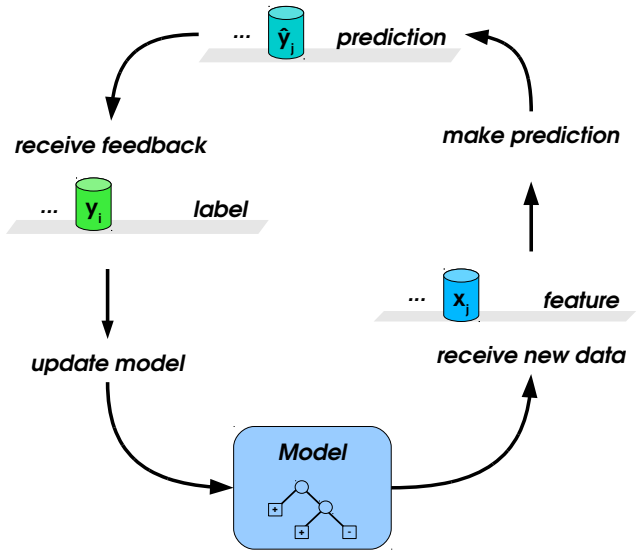
Concept Drift

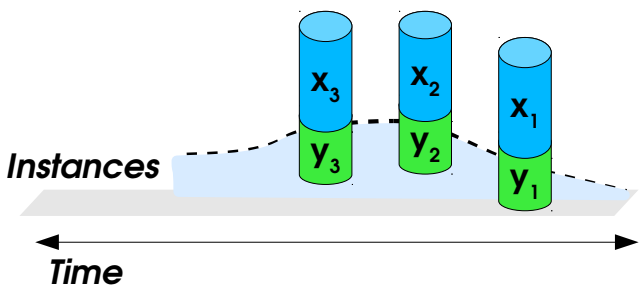
Denoted as

- *Concept drift*, e.g. in [Schlimmer and Granger, 1986]
- *Population drift*, e.g. in [Kelly et al., 1999]
- Related to *dataset shift* [Quiñonero-Candela et al., 2009]

Figure: Types of drift [Bifet et al., 2012], [Gama et al. 2013]

Classification in Streams





- Model update: *incrementally* or by model *replacement*
- But: Is that too simplified for real-world applications?

Figure: Typical Stream Classification Setting

Common Assumption:

Information (features, labels) on each instance is

- *correct* (i.e. reliable),
- *complete* (i.e. true labels and features finally known),
- *immediately available* (i.e. before the *next* instance must be processed)
- available at *no cost* and *without control* by the classifier *on label selection*.

Multiple Streams Notation

Formulation of stream classification using multiple streams

- Notation: Features x, true class labels y, predicted labels $\hat{y}$
- Related to scenario 2 in previous part (Spiliopoulou)
  - Instance = Entity
  - Every instance is only observed once
- Motivation
  - Distinction which information arrives when
  - Highlights *feedback issues*

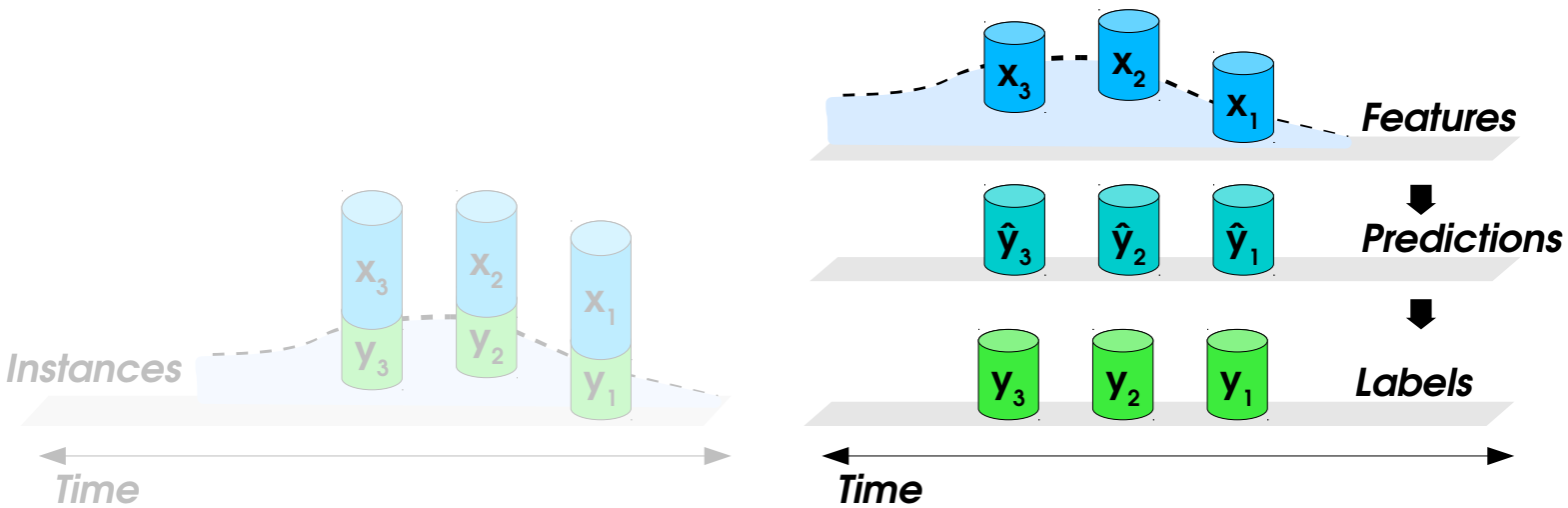


Figure: Typical stream classification problem (left) formulated using multiple streams (right)

Common Assumption:

Information (features, labels) on each instance is

- *correct* (i.e. reliable),
- *complete* (i.e. true labels and features finally known),
- *immediately available* (i.e. before the *next* instance must be processed)
- available at *no cost* and *without control* by the classifier *on label selection*.

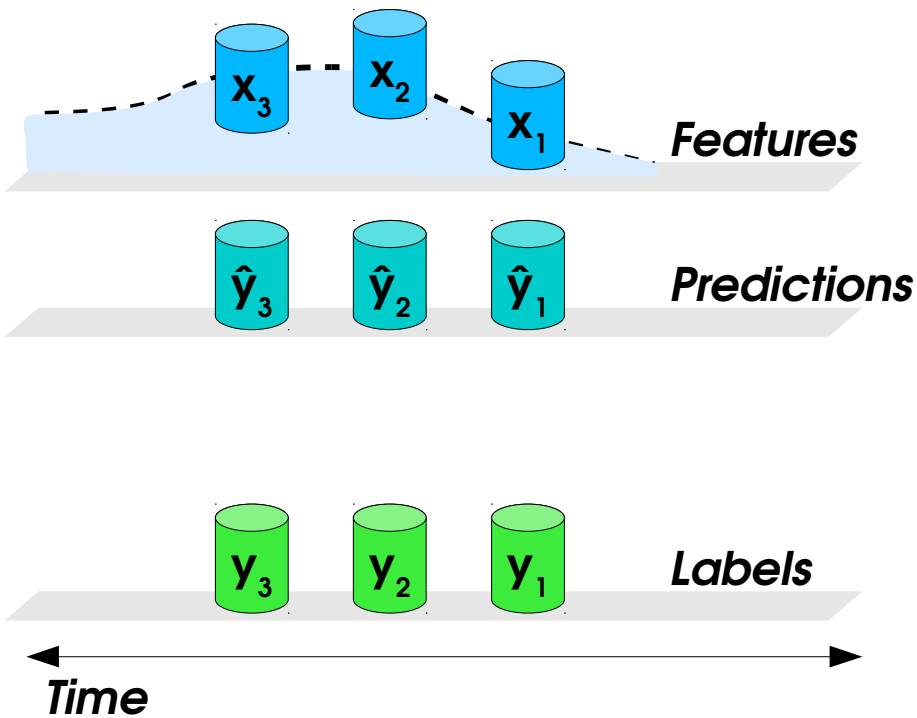


Figure: Typical Classification Setting in Streams

Label Paucity:

- *Completeness* of information is not met:
  Only some labels are available
- **Semi-Supervised Classification**

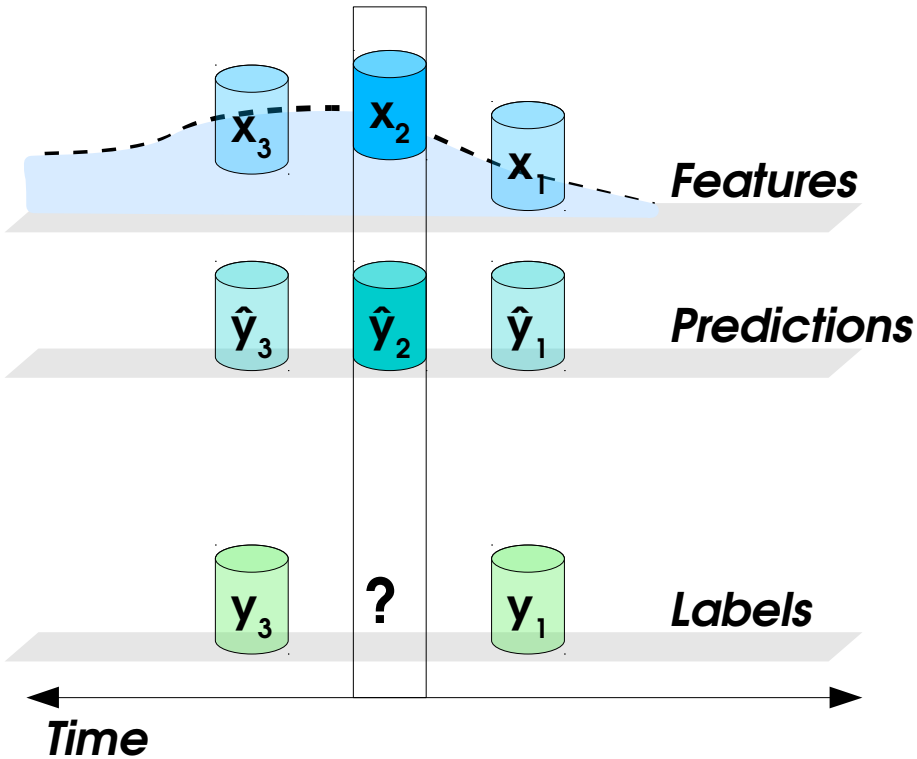


Figure: Label Paucity in Streams

Verification Latency:

Definition

Instantaneous availability of information is not met:
Labels arrive with delay

Motivating Examples

Whenever predictions concern *outcomes far in the future*, e.g.

- Credit scoring: Loans of long maturity
- Long-term stock market prediction

Types of latency:

- *Fixed* latency (label arrive with the same delay)
- *Varying* latency
  Labels arrive in different order than features
  - *Random* (not dependent on label)
  - *Dependent on label*:
    Class prior estimation?

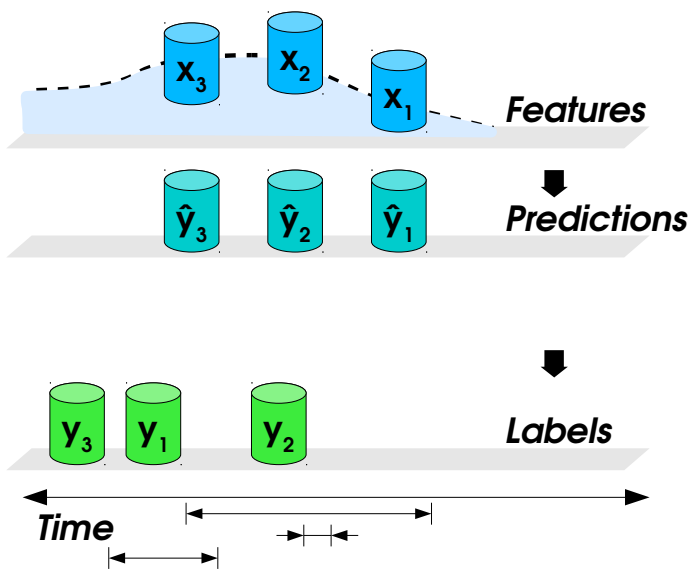
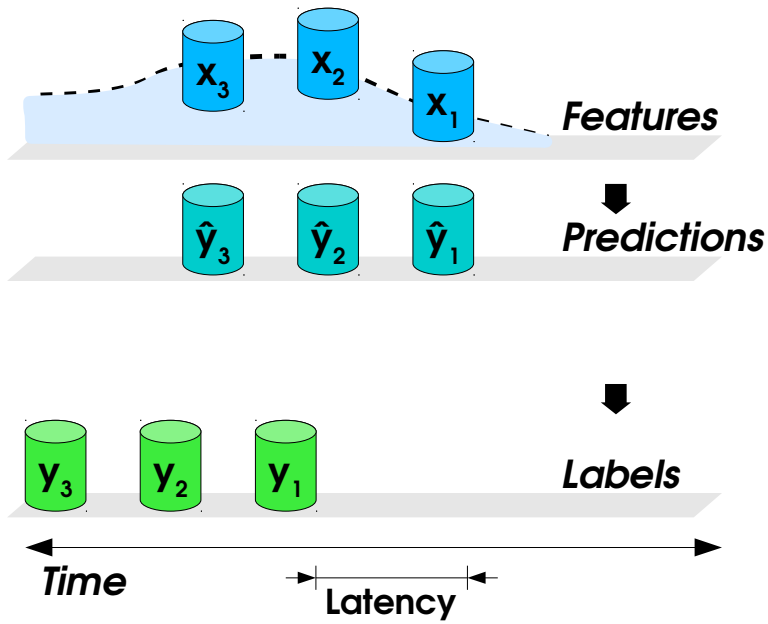


Figure: Latency in Streams

Verification Latency

Definition

- Latency between prediction and its verification
- Denoted as
  - *verification latency* [Marrs et al., 2010]
  - *label delay* [Kuncheva, 2008]
  - *time lag* [Lucas, 2004]
- Verification latency, no concept drift (similar to the *cold start problem*)
- *Concurrence* of *verification latency* and *concept drift*: **Major challenge**



Concurrence of Concept Drift and Verification Latency (CDVL)

Situation

Notation:

- Last label observed at time $t = 0$, last feature at time $t = 1$
- $P(\cdot)_t$ distribution at time t

- Drift can effect $P(Y|X)$, $P(X|Y)$, $P(Y)$, $P(X)$
- $P(X, Y)_0$, its derivatives and $P(X)_1$ are known[1],
- $P(X, Y)_1$ and thus $P(Y|X)_1$ are unknown
  - Labelled data corresponds to obsolete distributions, i.e. *no recent labelled data* is available
  - Available recent data is *unlabelled*

[1] More precisely, these distributions are directly assessable from the observed data.

Concurrence of Drift and Latency: Strategies

Possible Strategies

- Simply use of the most recent, labelled data?
  Problems:
  - Large latency and drift: **Obsolete prediction model!**
- Use available labelled and *unlabelled* information
  - Semi-supervised learning?
    No, features and labels are from different distributions!

Drift Detection

- [Zliobaitė, 2010] studies cases when detection is possible from *unlabelled* data:
- Possible if change in $P(X)$ is related to a change in $P(Y|X)$
- False positive alarms: Feature drift only ($P(X)$ changes, but $P(Y|X)$ not)
- False negatives: Posterior drift without feature drift
- Question: Can we turn this into adaptive *prediction* models?

Adaptive Prediction

- *Drift mining* approach [Hofer and Krempl, 2013]:
  - *Change mining* paradigm: "Understanding the changes themselves" [Böttcher et al., 2008]
  - Mining for relationships between drift of $P(X)$ and $P(Y|X)$:
    Formulate *drift models* about *temporal invariants* in this relationship
  - Requires historical, labelled data
  - Update prediction model using new, *unlabelled* data
  - Requires knowledge of drift model
  - Model can be updated on new instances *before* classifying them!
  - References:
    - [Krempl, 2011a] and [Hofer and Krempl, 2013] study models of drifting class priors
    - [Alaiz-Rodriguez et al., 2011] mixture model, drifting class prior and drifting mixing proportions
    - [Krempl and Hofer, 2011] and [Krempl, 2011b] drifting (non-)parametric mixture models
- Relationship to *transfer learning*?

Concurrence of Drift and Latency: Relationship to Transfer Learning

Transfer Learning (TL)

- Distinct source S and target T domains
- Full information in source domain: $P(X, Y)_S$ and its derivatives are known
- Partial information in target domain *Unsupervised transductive TL* or *Unsupervised Domain Adaptation*[1] $P(X)_T$ is known, $P(X, Y)_T$ is unknown
- Objective: *Knowledge transfer* from source to target domain, find *common hypothesis*

- (Some) similarity of tasks assumed[2]
- Any relationship, not necessarily time
- Shift-like relationship
- Moment of change is known [Ramon et al., 2007]
- Mostly batch-processing

Drift Mining

- Distributions drift over time
- Full info about past distributions: $P(X, Y)_0$ and its derivatives are known
- Partial info about current distributions: *Verification Latency*: No recent, labelled data $P(X)_1$ is known, $P(X, Y)_1$ is unknown
- Objective: *Adapt* classifier to new distr., identify *temporal invariants* in drift

- Posteriors are different (*real concept drift*)
- Explicit temporal relationship of S and T
- Gradual drift: Smooth transitions
- Sudden shift: Change point is unknown
- Online, incremental processing

But: Some synergies between TL and latency and drift mining problem settings, e.g. [Forman, 2006] on temporal inductive transfer for recurring context..

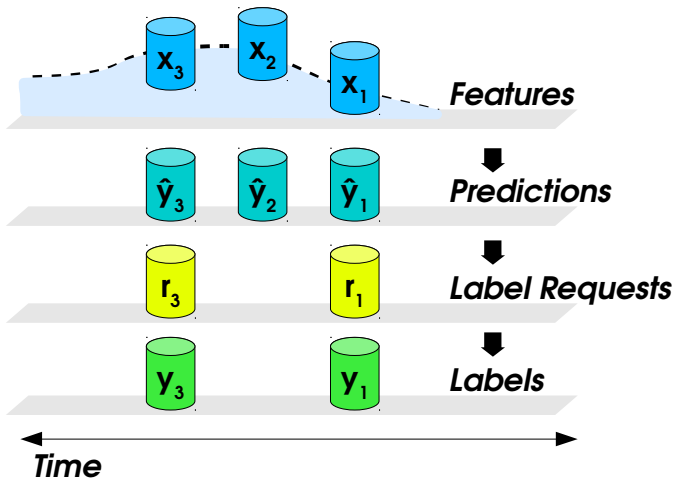
[1] See [Arnold et al., 2007] and [Jiang, 2008].

[2] Some definitions in the transfer learning literature are contradictory, see appendix.

Active Learning

Definition

- Label availability at *no cost* and *without selective control* not met:
  - Labels are costly (i.r.t. features)
  - Learner controls labelling process
- Objective: Strategy for selection of most valuable labels
- AL in streams with static concepts: well-studied, e.g. in surveys by
  - [Settles, 2009]: Section on *stream-based selective sampling*
  - [Fu et al., 2012]: Section on *AL on streaming data platform*
- Our focus: AL in *drifting* streams (selective sampling in evolving streams)



AL in Evolving Streams

Motivation

Why not simply apply active learning strategies from static (*iid*) streams?

- Drift can affect *any region* of feature space [Zliobaitė et al., 2011]
- Management of labelling budget (Convergence?)

Figure: Location of Drift[a]

[a] From [Zliobaitė et al., 2011], Figure 6, page 605.

AL in Evolving Streams: Self Lock-In Problem

Motivation

Why not simply apply active learning strategies from static (*iid*) streams?

- Example: Uncertainty sampling, *drifting* distributions
- Error is *never* even noticed!
- **Active learner (self) lock-in** on an outdated hypothesis
- **Caveat:** Drift can occur anywhere in the feature space, as noted by [Zliobaitė et al., 2011]
- **Remedy:** Sampling from the whole feature space.

AL for Evolving Streams: Selected Techniques (1)

Budget Management

Development of methods for estimating and controlling the labelling budget over time.

Motivation 1: Estimating the required labelling effort over time:

- Static context: Decreasing labelling efforts through convergence
- Dynamic context: Not necessarily the case...

Motivation 2: Balance of labelling costs over time:

- Simplistic approach: Random sampling of a fixed percentage
- What to do for more sophisticated AL strategies?

Relevant work:

- [Zliobaitė et al., 2011]
  - **Variable uncertainty:**
    Sampling the least certain instances in a window,
    and adjusting the window if drift is suspected.
  - **VU with randomisation:**
    As above, but include randomness for diversity over feature space.
- [Zhu et al., 2010]
  - **Minimum-variance approach** for estimating the number of required instances,
  - **Random sampling** for diversity over feature space.

AL for Evolving Streams: Selected Techniques (2)

Change Detection

Monitoring of the feature distribution for changes

Motivation: Unlabelled instances are cheap and their distribution is unbiased. Some changes in the feature distribution might hint to concept drift.

Advantage: Requires no labelled instances

Problems: Changes in posterior might go unnoticed (false negatives), can also trigger false alarms (covariate drift without concept drift).

Relevant work:

- [Fan et al., 2004] and [Huang and Dong, 2007] monitor changes in distributions of the leafs of a decision tree
- [Masud et al., 2010] use outlier detection to monitor changes in regions of previously low density.

AL for Evolving Streams: Selected Techniques (3)

De-Biasing

Use importance sampling to reweigh labelled data, as to feed an unbiased training set to the classifier.

Motivation: Distribution of labels is biased by the label selection process in active learning.

Challenge: Control of variance

Relevant work:

- [Chu et al., 2011] adopt importance sampling techniques to AL in concept drifting streams.
- Discuss the design of optimal instrumental distributions
- Compare performance of online Bayesian linear classifiers trained on biased and unbiased data

AL for Evolving Streams: Literature Overview

Reference	Stream Handling	Drift Type	Act. Learn. Strategy	Required Budget
[Fan et al., 2004]	online	feature	triggered Rand	fixed, on event
A change detector on $P(X)$ triggers random sampling, a predefined budget is spent upon change detection.				
[Huang and Dong, 2007]	chunks	feature	triggered US	fixed, on event
As above, but Naive Bayes-based uncertainty sampling.				
[Zhu et al., 2007]	chunks	any	MinVar QbC	fixed
Fixed proportion of a new chunk is labelled randomly and used to train a new classier, the ensemble variance is used for selecting upon the remaining instances. [Zhu et al., 2010] extends this work and determines required number of labels automatically.				
[Masud et al., 2010]	chunks	any	QbC, outlier	varying
Ensemble of pseudopoints (labelled clusters) is maintained, labels are requested for outliers outside all pseudopoint ranges and for instances with high disagreement (QbC).				
[Lindstrom et al., 2010]	chunks	posterior	US	fixed
The distance to hyperplane of SVM classifier is used for selection.				
[Liu and Wang, 2011]	online	posterior	US QbC	varying
Ensemble of field classifiers is maintained, for a new instance the ensemble variance is compared to the historical average.				
[Chu et al., 2011]	online	any	US	managed
Use uncertainty of linear probit model, but - model uncertainty is incorporated explicitly, - use importance weighting for **de-biasing**.				
[Zliobaitė et al., 2011]	online	any	US, Rnd	managed
Discuss problem of drift in arbitrary location of feature space, discuss several methods for **budget management**.				

Acronyms: Rnd = Random Sampling, US = Uncertainty Sampling, QbC = Query-by-Committee, MinVar = Minimum Variance

The theoretical work of [Yang, 2011] proves bounds for the number of mistakes and label requests in covariate drift (i.e. static concept).

Conclusion

Formulation Using Multiple Streams:

- Framework to highlight feedback issues such as:
  - Reliability
  - Completeness
  - Immediate availability: Verification Latency and Drift Mining
  - Labelling Cost and Selective Control: Active Learning

Verification Latency:

- *Challenging problem:* No recent labels at all!
- *Drift Mining:* Mine for invariants *in drift* over time
- Related to *transfer learning*, some synergies possible

Active Learning:

- Drift can occur anywhere in feature space!
- Budget management is non-trivial
- Only limited work on *bounds for errors and label requests* yet:
  For covariate drift without posterior drift: [Yang, 2011]

Selected References

Transfer Learning

- Comparative study on transductive TL methods: [Arnold et al., 2007]
- Survey on domain adaptation: [Jiang, 2008]
- Survey on transfer learning: [Pan and Yang, 2010]

Dataset shift

- Survey on dataset shift: [Moreno-Torres et al., 2012]
- Relation between dataset shift and transfer learning [Storkey, 2009] in [Quiñonero-Candela et al., 2009]

Active Learning

- Survey by [Fu et al., 2012] and [Settles, 2009]

Some Research Challenges in Mining Real-World Data Streams

- Dealing with realistic data and workflows
  - Availability and delay of feedback
  - Reliability / correctness of feedback
  - User participation to varying degrees
  - Interactive user feedback
  - Scalability
- Integrating expert knowledge
  - What to ask?
  - When to ask?
- Moving from adaptive algorithms towards adaptive tools
  - Adaptive pre-processing
  - Improving usability and trust
  - Autonomous systems, self-diagnosis

Closing and Outlook

Upcoming events:

- Workshop at ECMLPKDD 2013, September 23 in Prague, Czech Republic:
  Real-World Challenges for Data Stream Mining (RealStream)
  Organised by George Forman, Georg Krempl, Yin Wang, Indrė Zliobaitė.
  See `https://sites.google.com/site/realstream2013` for details.

- Session 6A at PAKDD 2013, tomorrow 14:00:
  Stream Data Mining, Chair Vladimir Estivill-Castro

- Position opening: Stream mining for medical research
  Opening details are at the reception desk, contact us in the lunch break.

Questions?

Thank you and enjoy your meal!

Bibliography

Alaiz-Rodriguez, R., Guerrero-Curieses, A., and Cid-Sueiro, J. (2011).

Class and subclass probability re-estimation to adapt a classifier in the presence of concept drift.
Neurocomputing, 74(16):2614–2623.

Arnold, A., Nallapati, R., and Cohen, W. W. (2007).

A comparative study of methods for transductive transfer learning.
In *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops*, ICDMW '07, pages 77–82, Washington, DC, USA. IEEE Computer Society.

Bifet, A., Gama, J., Gavaldá, R., Krempl, G., Pechenizkiy, M., Pfahringer, B., Spiliopoulou, M., and Zliobaitė, I. (2012).

Advanced topics on data stream mining.
Tutorial at the ECMLPKDD 2012, Bristol, UK.

Böttcher, M., Höppner, F., and Spiliopoulou, M. (2008).

On exploiting the power of time in data mining.
ACM SIGKDD Explorations Newsletter, 10(2):3–11.

Chu, W., Zinkevich, M., Li, L., Thomas, A., and Tseng, B. (2011).

Unbiased online active learning in data streams.

In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 195–203, New York, NY, USA. ACM.

Fan, W., Huang, Y.-a., Wang, H., and Yu, P. S. (2004).

Active mining of data streams.

In *Proceedings of the 4th SIAM International Conference on Data Mining, SDM 2004, USA*, pages 457—461.

Forman, G. (2006).

Tackling concept drift by temporal inductive transfer.
In *Proceedings of the ACM SIGIR Conference*, pages 252–259. ACM Press.

Fu, Y., Zhu, X., and Li, B. (2012).

A survey on instance selection for active learning.
Knowledge and Information Systems, 35(2):249–283.

Hofer, V. and Krempl, G. (2013).

Drift mining in data: A framework for addressing drift in classification.
Computational Statistics and Data Analysis, 57(1):377–391.

Huang, S. and Dong, Y. (2007).

An active learning system for mining time-changing data streams.
Intelligent Data Analysis, 11(4):401–419.

Jiang, J. (2008).

A literature survey on domain adaptation of statistical classifiers.
Technical report, Computer Science Department at University of Illinois at Urbana-Champaign.

Kelly, M. G., Hand, D. J., and Adams, N. M. (1999).

The impact of changing populations on classifier performance.
In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 367–371.

Krempl, G. (2011a).

Adaptive Prediction Models and their Application to Credit Scoring.
PhD thesis, University of Graz.

Krempl, G. (2011b).

The algorithm APT to classify in concurrence of latency and drift.
In Gama, J., Bradley, E., and Hollmén, J., editors, *Advances in Intelligent Data Analysis X*, volume 7014 of *Lecture Notes in Computer Science*, pages 222–233. Springer.

Krempl, G. and Hofer, V. (2011).

Classification in presence of drift and latency.

In Spiliopoulou, M., Wang, H., Cook, D., Pei, J., Wang, W., Zaïane, O., and Wu, X., editors, *Proceedings of the 11th IEEE International Conference on Data Mining Workshops (ICDMW 2011)*. IEEE.

Kuncheva, L. I. (2008).

Classifier ensembles for detecting concept change in streaming data: Overview and perspectives.
In Okun, O. and Valentini, G., editors, *Proceedings of the second workshop on supervised and unsupervised ensemble methods and their applications (SUEMA2008)*, volume 245 of *Studies in Computational Intelligence*, pages 5–10. Springer.

Lindstrom, P., Delany, S., and Mac Namee, B. (2010).

Handling concept drift in a text data stream constrained by high labelling cost.

In *Proceedings of the 23rd Int. Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)*, pages 32–37.

Liu, W. and Wang, T. (2011).

Online active multi-field learning for efficient email spam filtering.
Knowledge Information Systems.
Online First.

Lucas, A. (2004).

Updating scorecards: Removing the mystique.
In Thomas, L. C., Edelman, D. B., and Crook, J. N., editors, *Readings in Credit Scoring*, pages 93–110. Oxford University Press.

Marrs, G., Hickey, R., and Black, M. (2010).

The impact of latency on online classification learning with concept drift.
In Bi, Y. and Williams, M.-A., editors, *Knowledge Science, Engineering and Management*, volume 6291 of *Lecture Notes in Computer Science*, pages 459–469. Springer.

Masud, M. M., Gao, J., Khan, L., Han, J., and Thuraisingham, B. (2010).

Classification and novel class detection in data streams with active mining.
In *Proceedings of the 14th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining - Volume Part II*, PAKDD 2010, pages 311–324, Berlin, Heidelberg. Springer-Verlag.

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. (2012).

A unifying view on dataset shift in classification.
Pattern Recognition, 45(1):521–530.

Pan, S. J. and Yang, Q. (2010).

A survey on transfer learning.
Transactions on Knowledge and Data Engineering, 22(10):1345–1359.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D., editors (2009).

Dataset Shift in Machine Learning.
MIT Press.

Ramon, J., Driessens, K., and Croonenborghs, T. (2007).

Transfer learning in reinforcement learning problems through partial policy recycling.
In *European Conference on Machine Learning (ECML 2007)*, volume 4701 of *Lecture Notes in Computer Science*, pages 699–707. Springer.

Schlimmer, J. C. and Granger, R. H. (1986).

Beyond incremental processing: Tracking concept drift.
In *AAAI*, pages 502–507.

Settles, B. (2009).

Active learning literature survey.
Computer Sciences Technical Report 1648, University of Wisconsin-Madison, Madison, Wisconsin, USA.

Storkey, A. (2009).

When training and test sets are different: characterising learning transfer.
In *Dataset Shift in Machine Learning*, pages 1–28. MIT Press.

Yang, L. (2011).

Active learning with a drifting distribution.
Neural Information Processing Systems.

Zhu, X., Zhang, P., Lin, X., and Shi, Y. (2007).

Active learning from data streams.
In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 757–762, Washington, DC, USA. IEEE Computer Society.

Zhu, X., Zhang, P., Lin, X., and Shi, Y. (2010).

Active learning from stream data using optimal weight classifier ensemble.
IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 40(6):1607 – 1621.

Zliobaitė, I. (2010).

Change with delayed labeling: When is it detectable?
In *IEEE International Conference on Data Mining Workshops (ICDMW 2010)*, pages 843 – 850.

Zliobaitė, I., Bifet, A., Pfahringer, B., and Holmes, G. (2011).

Active learning with evolving streaming data.
In *Proceedings of the 21st European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD'11)*, volume 6913 of *Lecture Notes in Computer Science*, pages 597–612. Springer.